藏文分词及其在藏汉机器翻译中的应用

孙萌 华却才让 姜文斌 吕雅娟 刘群

摘要：本文提出一种基于判别式模型的藏文分词方法，并研究了藏文分词在藏汉机器翻译中的应用。根据藏文构词特性，通过最小构词粒度切分、感知机解码和分词结果重排序三个模块，显著提升了藏文分词质量。在此基础上，我们还提出了基于词图的藏汉机器翻译方法，缓解了分词错误在翻译中的传播，可以使翻译质量明显提高。

关键词：藏文 分词 机器翻译 词图

1 引言

藏文是一种具有逻辑格语法体系的拼音文字，表达方式和构词方式较为灵活，又因为受到周边国家或地区语言文化的影响，具有多种语言的特性。藏文分词技术是藏文信息处理的基础。在藏文中，和汉语类似，词与词之间都是字符或字的序列，缺乏间隔标记。因此藏文信息处理面临和汉语同样的问题，即如何将字符序列切分成合理的词语序列。而现有的藏文分词技术切分的准确率和实用性还有待提高。这一方面是因为藏文编码方案较多并且藏文研究起步较晚，另一方面更是由于藏文本身较为复杂的构词规律所致。

藏文信息处理领域中的信息检索、词法分析、句法分析、语义消歧、机器翻译、数据挖掘、舆情监控等应用研究的第一步就是分词。许多学者在藏文分词领域做出了很多研究成果，为进一步进行藏文语言处理相关技术的发展提供了必要的基础和开创性的思路。传统的藏文分词技术采用的是基于规则的方法，在考虑藏文特殊构词规律的基础上，用词典匹配实现藏文的初步分词。但是，基于规则的分词方法的分词效果较差。近年来，汉语分词技术快速发展，基于统计的分词模型在汉语分词上获得较大成功。然而，单独使用统计模型并不能很好地刻画藏文的构词特性，因此，将基于规则的方法和统计的方法融合是藏文分词技术的发展方向。

机器翻译是指用计算机将一种语言翻译为另一种语言。机器翻译技术致力于解决不同语言之间的交流障碍。藏族文化源远流长，藏文承载了丰富的文化遗产，在民族交流日益频繁的今天，机器翻译技术将有效地把藏族的文化和智慧转为用汉语表达，以促进藏族文化的传播和发展。现有的藏汉机器翻译系统采用基于规则的方法，即词典和人工制定的翻译规则相结合的方法。基于规则的方法的劣势在于，制定规则的人工成本过高，可移植性差。对于通用的机器翻译技术而言，基于统计的方法已经成为当今的主流。基于统计的机器翻译的优势在于，仅需要构建双语平行语料库，不需要很多人工的介入，统计模型就可以从中自动学习到翻译知识，实现双语之间的翻译。构建双语平行语料库的人工成本远小于人工制定翻译规则的成本，并且平行语料库的规模越大，其训练出来的翻译模型的翻译性能越高。

随着信息科技的迅速发展，藏语信息处理技术也取得了极大的进展，但是由于起步较晚，还处于较为初级的阶段。将面向藏文处理的规则方法和统计方法进行有机的结合，会推进藏文信息技术的进一步实用化。

2 藏文构词特点及藏文分词研究现状

2.1 藏文构词简介

藏语和汉语同属于汉藏语系，具有一些相同的特点：(1). 汉语和藏语都是单音节构字，即一个字中只有一个元音；(2). 都具有量词；(3). 以虚词或语序作为表达语法意义的重要途径；(4). 词与词之间没有空格。汉语与藏语的最大不同在于其文字的表现形式，汉语是字符文字，而藏语是拼音文字。因而藏文句子的书写表现形式类似于汉语的拼音形式的拼接，值得注意的是，音节之间没有空格分割。

藏文的拼音字母是由 4 个元音和 30 个辅音组成，称之为构件。藏文各音节之间用一个“点”分隔，称之为音节点。一个音节最多由 7 个构件组成，但其组合方式包括水平和垂直组合。选择一个辅音作为基础构件，在其上下左右根据一定规律放置其它构件，比如，元音往往会置于上方或下方。

一般而言，藏文的音节是书写的最小单元，可以理解为汉语中的一个字，为了便于表达，本文也将藏文的音节称为字。如图 1 所示：



图1. 藏文词“教室”

前面的一个音节表示“学习”的意思，后面的一个音节表示“房间，屋子”的意思。在藏文中大量地使用格助词和紧缩词。有些格助词，如图 2 所示，在特定语境下可以省略前面音节的音节点，直接拼接在前一个音节之后。由于这种紧缩现象的频繁出现，不能简单地将音节作为藏文分词的基本粒度，因此，藏文的基本粒度切分是藏文分词的第一步。

藏文被转换为分词的基本粒度之后，可以借鉴汉语自动分词方法。这里的分词基本粒度在语言学上并不具有任何意义。基本粒度往往比一个音节要小，我们可以称基本粒度为分词所用的“字”，转换为字序列之后，我们再进行相关的分词研究。而字序列的分词，就可以借鉴汉语中已经成熟的方法。

ར　ས　འི　འོ

图2. 藏文格助词

2.2 藏文分词相关工作

现有的藏文分词方法大体可以分为两类。第一类是基于藏文特有的语言学知识的规则方法。陈玉忠[1][2]从藏文的字切分特征、词切分特征和句切分特征三个方面深入研究藏文特有的语法接续规则，提出了基于格助词和接续特征的藏文分词方法。才智杰[3][4]首先识别格助词，然后将其作为分隔符对句子进行切分，采用最大匹配的算法依次对切分之后的“块”进行分词。基于规则的藏文分词算法通常需要一个规模足够大的词典，采用最大匹配算法，即以在词典中查到的最长的词条作为句子的切分，算法实现简单，效率较高。这种方法不能很好地处理切分歧义问题，另外对于未登录词的识别能力不强。第二类是基于统计的机器学习方法，用到的统计模型主要是隐马尔科夫模型[5][6]。然而相对于复杂的藏文构词现象，隐马尔科夫模型仍然略显简单。当前，其他统计模型，如最大熵[7]、感知机和条件随机场等判别式模型的应用已经成为汉语分词方法的主流方向。判别式模型的优势在于同时支持简单的特征和复杂的特征，特征空间具有回退特性，因此具有较好的模型泛化能力。可以预见这些模型也应能够用于藏语。

3 基于判别式模型的藏文分词方法

3.1 藏文分词系统

针对藏文分词过程各个层面的处理对象以及问题特点，我们的藏文分词系统包含了原

子切分、基于感知机解码和分词结果重排序三个主要模块，系统的流程如图 3 所示。

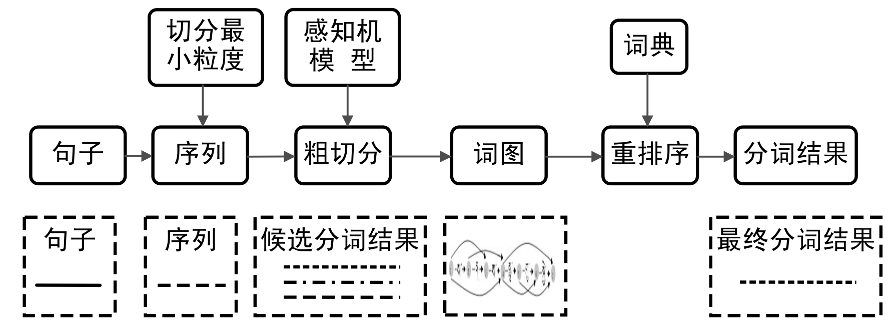


图3. 藏文分词系统流程图

首先，根据藏文特有的构词规律将句子切分成最小粒度的序列 — 单元序列；随后，根据感知机模型提供的判别式分类的权重，在单元序列上进行维特比解码，从而生成有向图，并通过查询词典为各条边赋予不同的权重；最后，通过最短路径算法求解加权有向图中的最短路径，生成最终分词结果。

下面对这几个步骤分别进行介绍。

3.2 藏文最小构词粒度切分

一个或多个藏文字丁组成一个音节，一个或者多个音节组成词，音节之间由音节点分隔。如图 4 所示的藏文片段，其汉语含义是：处在某一个级别。

基于序列标注模型的汉语分词过程可以视为在字层面上的组合。而藏文分词过程的复杂性在于，不能直接在音节层面上进行组合，在有些情况下需要将某个音节拆分，和左边或者右边音节组合，或者独立成词。由于这种组合的灵活性，对于藏文的标注序列的最小构词粒度必须是小于音节的单位。我们设计了三种构词粒度的方案以描述其构词规律，如图 5 所示：

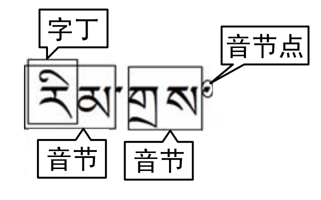


图4. 藏文片段

方案 1：**以藏文字丁为构词粒度** 对句子按照每个字丁进行切分，如图 5 中的切分（a）。

方案 2：**以藏文字丁－音节点为构词粒度** 不将音节点单独切分出来，而是将其与左边字丁组合，如图 5 中的切分（b）。

方案 3：**以音节为构词粒度** 定义特殊格助词表，先按照音节扫描切分句子，一旦音节中含有特殊格助词则匹配相对应的规则切分此音节，如图 5 中（c）所示。

选择藏文的最小构词粒度的关键在于将藏文句子切分为基本粒度序列。基本粒度指的是无需进一步切分的“字”，而分词过程可以看做是连续的基本粒度的组合，进而成词。（a）方案没有考虑任何构词规律，在分词标注语料有限而字丁构词现象又较为复杂的情况下，统计模型缺乏足够的知识进行标注学习；（b）方案在（a）方案的基础上考虑音节规律，减小了分词时的解码搜索空间；而方案（c）则最大程度保存了音节的内部结构，却又不会破坏构词粒度的原子性。然而较大的粒度在规模受限的标注语料中会出现

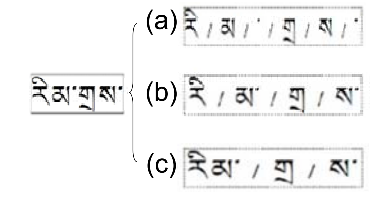


图5. 三种构词粒度切分方案举例

数据稀疏的问题。实验部分探讨了在同样规模标注语料库下，采用以上不同的切分策略对最终分词效果的影响。

3.3 基于感知机模型的藏文分词方法

感知机是线性判别式模型，形式简单，根据特定任务设计出合适的特征，会收到非常显著的分类效果。传统的感知机用以解决两分类问题，如果模型计算一个实例的特征向量得分大于某个阈值，此实例属于+1 类，否则属于-1 类。但是，自然语言处理任务中更常遇到的是多分类（不止两类）问题。对于分词任务而言，需要判断每个字的所属分类，根据字分类情况，产生分词结果。通常定义 4 种类别，分别是：

– b: 词的开始
– m: 词的内部
– e: 词的结尾
– s: 字单独成词

为解决上述问题，我们可以将其进行转化。对于每个字，通过模型分别计算其在属于 4 种类别时的模型得分，选择最高分的类别作为这个字的最终归属类别。但是分词过程并不是单独的对字进行分类，还要考虑到相邻字分类类别的兼容性。由上面对类别的定义，我们可以推导出下面的规则。

1. 如果当前字的类别是 b，其后面字的类别不能是 b；
2. 如果当前字的类别是 m，其后面字的类别只能是 m 或者 e；
3. 如果当前字的类别是 e，其后面字的类别只能是 b 或者 s；
4. 如果当前字的类别是 s，其后面字的类别只能是 b 或者 s。

我们可以通过维特比算法对基本字序列进行序列标注，而序列标注的权重由感知机模型训练得到，其中感知机模型是判别式分类模型的一种。柯林（Collin）[8]提出的基于感知机的序列标注方法是一种在线的学习方法，将传统的感知机训练算法应用于分词任务，训练时对于正确标注增加其权重，对于错误标注减少其权重。感知机模型的训练速度快，分类效果好。我们采用平均感知机算法进行句子的粗切分，该算法记录每一次权重的改变，以提高分词系统的稳定性，算法如下所示：

平均感知机算法

input 训练实例(X,Y)	X，Y 构成平行语料库（共 N 对）
$W \leftarrow 0$；$j \leftarrow 0$；$W^j \leftarrow 0$	
for t=1 **till** T **do**	T 轮迭代
for i=1 **till** N **do**	
$z_i \leftarrow \arg\max_{z \in GEN(x)} \Phi(x_i, z) \cdot W^j$	选取得分最高标注序列，其中 z 为由 GEN（x）产生的标注结果。
$W^{j+1} \leftarrow W^j + \Phi(x_i, y_i) - \Phi(x_i, z_i)$	
$j \leftarrow j+1$	
end for	
end for	
output $W \leftarrow \sum_{k=1}^{j} W^k$	

设输入句子的原子序列 $x_i \in X$，输出标注序列 $y_i \in Y$, X 表示训练语料中的所有句子，Y

表示对应的标注，共有 N 个句子，其中{b, m, e, s}是标注的符号集合。

我们用函数 **GEN**(x)表示采用维特比算法产生输入句子 x_i 的候选标注结果，$\Phi(x_i, z)$ 表示输入句子和产生标注序列的特征向量，选择使得 $\Phi(x_i, z) \cdot W^j$ 得分最高的 z 作为标注序列。y_i 表示正确的标注序列，我们用正确标注序列的特征向量和产生的最好标注序列的特征向量之差更新权重 W，即仅更新被标记错误的字所对应的特征，增加这个字标准的类别的特征对应的权重，而减小字被错误标注的类别的特征对应的权重。

下面是一个例子。

如图 6 所示，这句藏文的意思是"苗从地里发芽"。假设我们只用特征模板

$$C_nC_{n+1}(n=-1..0)$$

首先由算法 1 中的第 5 条语句生成图 6 中的最后一个序列标注，这是一个错误的标注，第四个藏文字应该属于 e 类，却被预测为 s。对于标准序列在第四个藏文字上生成的特征是 feat1 和 feat2。

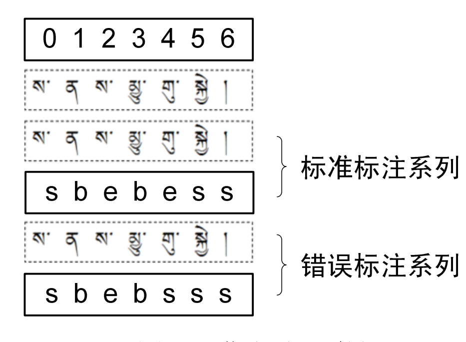


图6. 藏文分词举例

feat1: C−1C0=མྱུ་གུ་&&e

feat2: C0C1=གུ་སྐྱེ&&e

对于模型生成的错误的序列标注在第四个藏文字上生成的特征 feat3 和 feat4 是：

feat3: C−1C0=མྱུ་གུ་&&s

feat4: C0C1=གུ་སྐྱེ&&s

其中 C-1C0 对应前述特征模板表示 $C_nC_{n+1}(n=-1..0)$的 n，表示位置信息。&&表示分隔符，前面的藏文是文本特征，而后面的 e 是标注特征。算法描述中第六行指令的操作就是，将 feat1 和 feat2 对应的权重进行奖赏，而对 feat3 和 feat4 对应的权重进行惩罚。

3.4 分词特征设计

特征设计是在判别式训练中最为重要的任务，结果直接影响分词的质量。在设计特征时，需要研究者根据不同的任务需求，考察任务的特点，才能设计出合理的特征模板。

在分词任务中，对于当前字的分类，我们需要考虑到这个字前后相邻的字对其影响，所以要抽取当前字相邻的字作为特征。藏文词通常由较多的字组成，我们将特征模板的窗口设为 4，以抽取较多的特征刻画当前字。

特征模板，如表 1 所示。

其中，C_0 表示当前字，当前字左边的为 C_{-n}，如左边第一个字为 C_{-1}，同理当前字右边的字用 C_n 表示。

表1. 特征模板

$C_n\ (n=-4..4)$	$C_nC_{n+1}(n=-4..3)$	$C_{-2}C_{-1}C_1C_2$
$C_0C_n\ (n=-4..4)$	$C_0C_nC_{n+1}(n=-4..3)$	$C_0C_{-2}C_{-1}C_1C_2$

3.5 基于词图的重排序

对于基于感知机的分词，除了通常使用的局部特征，非局部特征的引入也会提升分词的性能。但是，非局部特征要在解码的过程中动态地生成，很难直接将其加入到分类器中，并且引入非局部特征也会影响训练过程中对应特征的调节。在自然语言处理的其他领域也面临着类似的问题，一般的解决方案是使用重排序的方法引入非局部特征。然而，传统的重排序是通过产生 *n*-best[1]结果。*n*-best 所能表示的搜索空间较小，并且储存了冗余数据。据此，我们在采用感知机模型进行藏文粗切分的过程中保存分词的候选，并将候选分词结果压缩为词图，最后采用基于词图的重排序算法寻找最合适的分词结果。

基于判别式模型的分词方法优势在于较高的泛化能力，用丰富全面的特征空间刻画分词结果，不论是未出现词还是已出现词，均可通过模型计算给出概率得分。但是，判别式模型的泛化能力可能导致常用词不能正确的切分，产生低级错误。因此，判别式模型和词典的有机结合会在一定程度上提高分词质量。

可以将词图看作一个有向无环图，如图 7 所示。以构词单元之间的空隙作为图的顶点，顶点之间的连线表示顶点之间的字符组合成词。每条边通过词典获得相应的权重，在词图上寻找一条最短路径，例如：<1, 4>、<4, 7>。

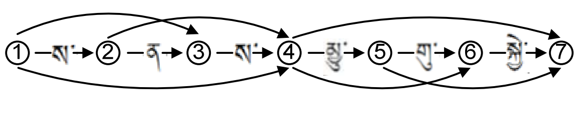


图7. 词图

按照词图顶点的拓扑顺序，对每个结点保存以此结点为终点的所有边，可以生成词图。例如，对于顶点 3，保存边<2, 3>和<1, 3>；对于顶点 4，保存边<1, 4>、<2, 4>和<3, 4>。

如果保存解码过程中的所有边，则词图中会包含过多的无用路径，在一定程度上会影响最短路径的生成。所以我们通过限制每个结点的入度，只保存得分最高的 n 条边，实现词图的简单剪枝。由于词图结构的特点，即使限制结点入度，同样会包含很多路径信息。

传统的最短路径分词原则是使切分出来的词数最少。在此基础上，加入词典惩罚特征，对每一条边通过查询词典赋予一个权重，通过动态规划算法求出最短路径。其中词典包含 95971 个常用词条。例如，对于藏文句子：

ས་ནས་མྱུ་གུ་སྐྱེ།　　（苗从地里发芽）

可能有如下三种切分：

1. ས་ / ནས་ / མྱུ་གུ་ / སྐྱེ / །
2. ས་ན / ས་མྱུ་ / གུ་ / སྐྱེ / །
3. ས་ / ནས་ / མྱུ་ / གུ་སྐྱེ / །

如果每条边的权重都为 1，则三种切分的分值都为 5。如果对于没有在词典中出现的词加以适当的惩罚，比如，将其边的权重调整为 2。切分方案 1 中，所有的词都在词典中存在，得分为 5；在切分方案 2 中，ས་མྱུ 和གུ 不会在词典中出现，得分为 1+2+2+1+1=7；切分方案 3，མྱུ 和གུ་སྐྱེ 不存在于词典中，得分为 1+1+2+2+1=7。因而切分 1 具有最短路径。

4　基于词图的藏汉机器翻译

藏汉机器翻译研究处于起步阶段，国内外还没有较为成熟的研究成果。主要原因在于：1. 藏文相关的信息处理基础性技术还没有达到实用阶段；2. 藏汉双语平行语料匮乏；3. 藏

[1] 前 n 个最佳结果

语语言学规律较为复杂，制定人工翻译规则的成本过大。目前，统计机器翻译是机器翻译的主流方向。统计机器翻译在近年来快速发展，翻译模型在众多研究者的努力下快速迭代发展，从最初的基于词的翻译模型，发展到基于短语的翻译模型，翻译质量得到极大的提升。但是，短语模型的翻译调序能力有限，难以实现翻译的长距离调序。蒋伟（David Chiang）[9]在短语模型的基础上创造性地提出了层次短语模型，从双语语料中自动抽取带有泛化变量的翻译模板，改进了翻译模型的翻译调序能力。层次短语模型可以比较好地完成两种语序差异较大语言之间的翻译，较好地解决了翻译过程中的长距离调序问题。汉语在语法上属于主动宾（主语、动词、宾语，SVO)结构，而藏文是主宾动（SOV）结构，在藏汉翻译中，层次短语模型具有较大的优势。

统计机器翻译在解码时通常需要两个步骤。第一步是分词、词形还原或者形态分析。根据语言的不同，选择不同的操作。汉语，藏文和泰语等词与词之间没有明显的分隔标记，需要的是分词处理；对于英语、德语等具有形态变化的语言，通常需要对其进行词形还原；有些形态丰富语言，譬如，维吾尔语、蒙语和芬兰语等，则需要形态分析。对于需要词形还原或形态分析的语言而言，直接使用单词原型通常会导致数据稀疏问题。如果将汉语或者藏文的字或者音节作为最小翻译粒度，由于短语模型或者层次短语模型在翻译时要求先划分短语，这就会导致意群片段的错误划分，从而使翻译质量会受到较大影响。第二步将第一步生成的词序列作为输入进行翻译解码。

对于藏汉翻译而言，分词的结果将直接影响下一步的翻译质量。首先，面对现实语料，分词器不可能实现完全正确的分词，而分词的错误在进入翻译模块时会导致错误的进一步传播，因此得到的译文也是不正确的。其次，较好的分词结果并不保证有较高的翻译质量，机器翻译是一个复杂动态的过程，因此分词的粒度应该由翻译模型决定。我们采用基于词图的层次短语翻译方法，显著提高了藏汉翻译的质量。

基于词图的统计机器翻译的输入是由分词产生的词图，如图 8 所示。

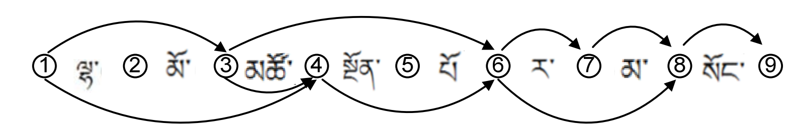


图8. 词图

这个藏文句子的意思是“毛拉没去青海湖”，“毛拉”是一个藏族的人名。图中的边所覆盖的字可以组合成词。这句话可以有多种分词方案，如果只将一种分词结果输入到翻译解码器中，翻译结果很可能出错。图 8 中的句子，不同的分词对应的目标端规则见表 2

表2. 翻译规则

藏语	汉语
ལྷ་མོ་	拉毛
ལྷ་མོ་མཚོ་	拉毛措
མཚོ་	湖
མཚོ་སྔོན་པོ་	青海湖
སྔོན་པོ་	蓝色
ར་མ་	山羊
མ་	没
སོང་	去

如果生成词图的边过少，并不能明显减少由于分词错误而导致的翻译错误；词图边过多，一方面增加解码器翻译的时间，另一方面无用的边或者错误的边会干扰正常的翻译。词图的生成是翻译的第一步，构建词图的质量将直接影响翻译的质量。我们采用两种剪枝策略：一是最大队列长度，二是最小入队权重。当生成边的权重小于最小入队权重，删除此边。如果将新生成的边放入队列后，长度超过最大队列长度，删除最小权重的边。

基于词图的解码将多种分词候选压缩为词图表示，分词结果对应词图中的一条路径，不

同路径可共享子路径，而基于词图的翻译的对象是词图中的边，不是某一种分词路径，因此不同路径共享的子路径只被翻译一次，这样可以减少冗余操作，加快解码的速度。

5 藏文分词方法和藏汉翻译实验

5.1 藏文分词实验

我们使用由青海师范大学提供的 12942 条人工分词的藏文句子，共包含 110K 词语，语料的领域较为广泛。从中随机选择 500 句作为测试集，剩余的作为训练集。

为了研究构词粒度对于基于感知机的藏文分词性能的影响，我们以基本字丁、基本字丁一音节点和音节为切分单位，设计 3 组实验，实验结果如表 3 所示。

我们发现，随着构词粒度的增大，分词结果的性能也在提升，基于音节的感知机藏文分词系统的 ***F*** 值比基于基本字丁的系统提高了 3.3 个百分点。可以将藏文分词看作序列标注的过程。增大构词粒度，则序列变短，分类器决策的次数将减少，减少了搜索空间，准确率提高。

表3. 藏文分词系统的性能

方法	准确率 (%)	召回率 (%)	***F*** 值 (%)
基本字丁	87.32	88.51	87.91
基本字丁一音节点	88.42	89.80	89.11
音节	91.21	91.22	91.21
音节+词图	**95.70**	**96.81**	**96.25**
基于规则（基线）	95.05	94.69	94.8

我们将以上三组实验看作是基于感知机模型的粗切分，分词结果中往往会出现不成词的切分。我们在基于音节的分词系统上加入基于词图的重排序模块，通过查询词典赋予每条边不同的权重，搜索出最短路径。最终分词的 ***F*** 值达到 96.25%，比基于规则的分词系统提高了 1.38 个百分点，比基于音节的分词系统[4]提高了 5.04 个百分点。

5.2 藏汉翻译实验

我们以层次短语系统作为基线系统，我们的系统在此基础上增加基于词图的规则匹配和解码功能。采用 SRILM[10]工具训练得到 5 元语言模型，以 Kneser Ney 方法进行平滑，其中训练的语料是 GIGAWORD 语料中的部分汉语单语语料，共 6.4 千万个字，使用中科院计算所开发的 ICTCLAS[11]对语料进行分词。

实验中使用主流的翻译评测指标 BLEU[12]，BLEU 通过统计翻译结果和参考译文之间的 N-Gram[2]匹配的准确率的几何平均衡量翻译质量。层次短语模型作为我们实验的基线系统，其输入的开发测试集采用的是藏文分词工具输出的 1-best 结果，词图解码的输入是由藏文分词工具输出的词图。

实验使用两组语料，如表 4 所示，实验 1 的训练语料主要来自政府公文和法律文献，开发测试集的领域也是政府公文相关的材料，语料相对比较正式。实验 2 的训练语料和开发测试集由政府公文、法律文献和日常口语组成，语料（特别是在口语领域）规范性较差，但环境更贴近现实情形。

表4. 两组藏汉实验的数据规模

	句数		
	训练集	开发集	测试集
实验数据 1	101629	650	517
实验数据 2	326698	1259	1096

[2] 大词汇语言处理常用的一种语言模型，该模型基于这样一个假设：第 n 个词的出现只与前面 N-1 个词相关，常用的是二元的 Bi-Gram 和三元的 Tri-Gram

互联网上的大部分藏文语料的随意性比较强，可能存在表达不规范，词语变形，词语缩略甚至错误的表达方式，这就要求翻译模型具有较高的容错性和鲁棒性，基于词图的翻译方法在一定程度上有助于此类问题的解决。

实验结果如表 5 所示，

表5. 实验结果（BLEU 得分）

	实验数据 1		实验数据 2	
	开发集	测试集	开发集	测试集
层次短语翻译模型	0.4201	0.3331	0.4901	0.3361
词图翻译模型	0.4314	0.3422	0.5080	0.3500

对于实验数据 1，词图翻译模型比层次短语模型在开发集上提高了 1.13 个百分点，在测试集上提高了 0.91 百分个点。传统的翻译模型以词的序列作为输入，但是错误的分词结果或者不合适的分词粒度都会影响翻译的结果。对比两个系统的翻译结果，词图翻译的结果明显减少了未登录词出现的次数。实验数据 2 的语料分布较为广泛，句子书写较为随意，比数据 1 的语料更接近日常生活用语，因此对分词工具和机器翻译模型的容错性的要求更高。词图翻译模型比层次短语模型在开发集上高出 1.79 个点，在测试集上高 1.39 个点。表明词图翻译模型的鲁棒性更强，更适合处理不规范文本。词图翻译不仅降低了未登录词的出现次数，还提高翻译的质量。

6 总结与展望

本文提出一种基于判别式模型的藏文分词方法，并探索构词粒度的大小对分词性能的影响，确定藏文分词的基本切分粒度。然而，由于非局部特征不能直接用于感知机，我们采用基于词图的重排序算法引入非局部特征，并运用最短路径算法产生最优的分词结果。分词过程或者形态分析是对汉语、藏文、泰文、维吾尔文或者朝鲜文进行机器翻译第一步。然而，输入翻译模型的分词序列，错误的分词或者不合适的分词粒度，都会导致翻译的错误。我们提出的基于词图的翻译模型，将多种分词结果压缩为词图表示，并作为机器翻译的输入，由翻译模型在翻译过程中选择最合适的分词粒度，提高了翻译质量。

在基于统计的汉语分词领域，最大熵和条件随机场也获得了较好的分词效果，未来我们将对比在同样特征集合下，哪个模型更适于藏文分词。另外，在重排序中我们仅考虑当前词的词典特征，没有考虑到当前词的上下文信息。下一步的工作将研究语言模型在重排序中的作用。藏文句法结构与汉语相差较大，在藏语端引入特定的规则，可以更好地指导藏汉统计机器翻译。

参考文献：

[1] 陈玉忠，李保利，俞士汶等. 基于格助词和接续特征的藏文自动分词方案. 语言文字应用，75-82. 2003

[2] 陈玉忠，李保利，俞士汶. 藏文自动分词系统的设计与实现.中文信息学报, 15-20. 2003

[3] 才智杰. 藏文自动分词系统中紧缩词的识别. 中文信息学报, 35-37. 2009

[4] 才智杰. 班智达藏文自动分词系统的设计与实现. 青海师范大学民族师范学报， 75-77. 2010

[5] 苏俊峰，祁坤钰，本太. 基于 HMM 的藏语语料库词性自动标注研究. 西北民族大学学报 , 42-45. 2009

[6] 史晓东，卢亚军. 央金藏文分词系统. 中文信息学报， 54-56. 2011

[7] Nianwen Xue, Libin Shen. Chinese word segmentation as LMR tagging . In Proceedins of the 2nd SIGHAN Workshop on Chinese Language Processing, in conjunction with ACL'03: 176-179

[8] Collins,Michael. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In Proceedings of the Empirical Methods in Natural Language processing Conference 2002: 1-8.

[9] David Chiang. Hierarchical Phrase-based Translation. Computational Linguistics, 2007, 33:201-228

[10] A.Stolcke. SRILM -- An Extensible Language Modeling Toolkit. In Proc. Intl. Conf. on Spoken Language Processing 2002: 901-904.

[11] Huaping Zhang, Qun Liu, Xueqi Cheng, and Hongkui Yu. Chinese lexical analysis using hierarchical hidden markov model. In Proceedings of the second SIGHAN workshop on Chinese language processing 2003 : pages 63–70.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of ACL 2002 : 311–318.

作者简介：

孙 萌： 中科院计算所智能信息处理重点实验室、硕士研究生 sunmeng@ict.ac.cn

华却才让： 青海师范大学、副教授

姜 文 斌： 中科院计算所智能信息处理重点实验室、助理研究员

吕 雅 娟： 中科院计算所智能信息处理重点实验室、副研究员

刘 群： 中科院计算所智能信息处理重点实验室、研究员